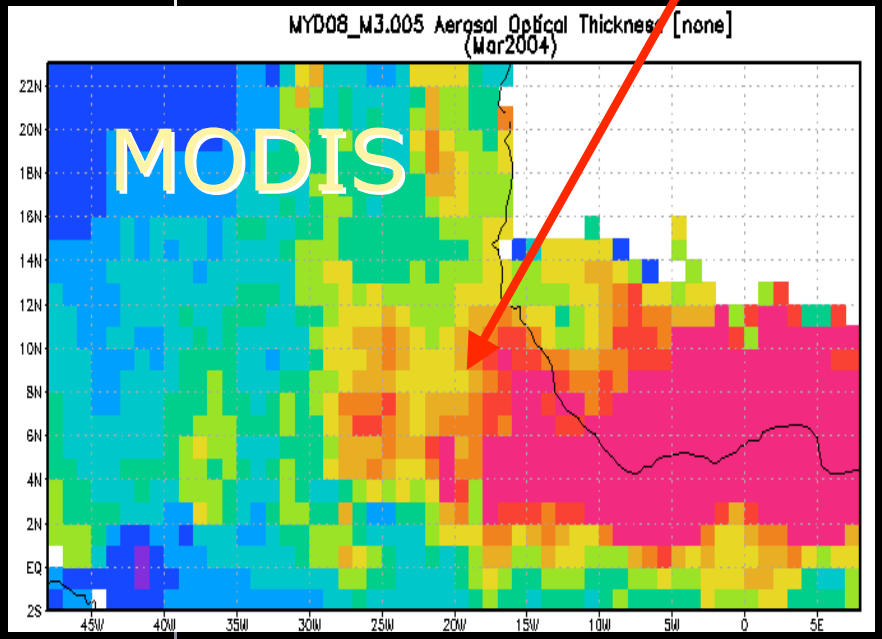
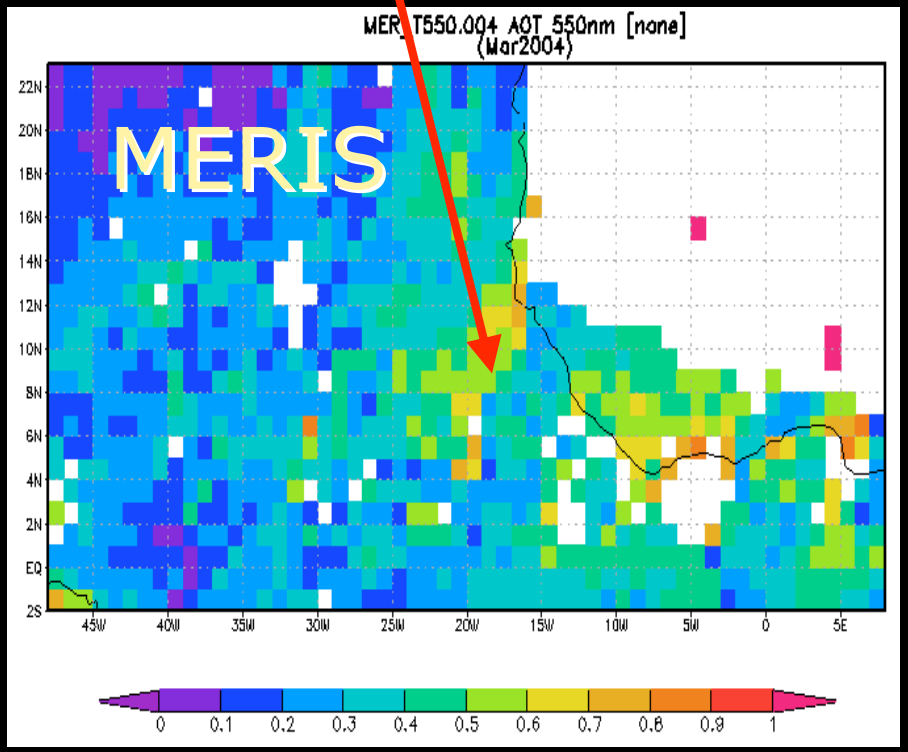
Science Quality and Data provenance

Data Provenance: *the source of data, including the execution history of the processes that produced them*

Same parameter

Same space & time





Different provenance → Different result

Data Provenance and Science Quality

- We can save time by providing data, tools and services to scientists but…
- Science quality of our products is imperative for scientists to be able actually trust and use them
- Documenting all the steps leading to the final product is paramount
- Also, providing assessment of sensitivity of the results to variations in processing algorithms/steps… published in peer-reviewed papers and presented to users in convenient, easy-to-find-and-read fashion
- Only working closely with scientists can guarantee science quality
- Scientific credentials (peer-reviewed papers on the subject) help to produce scientifically correct results

Problem definition

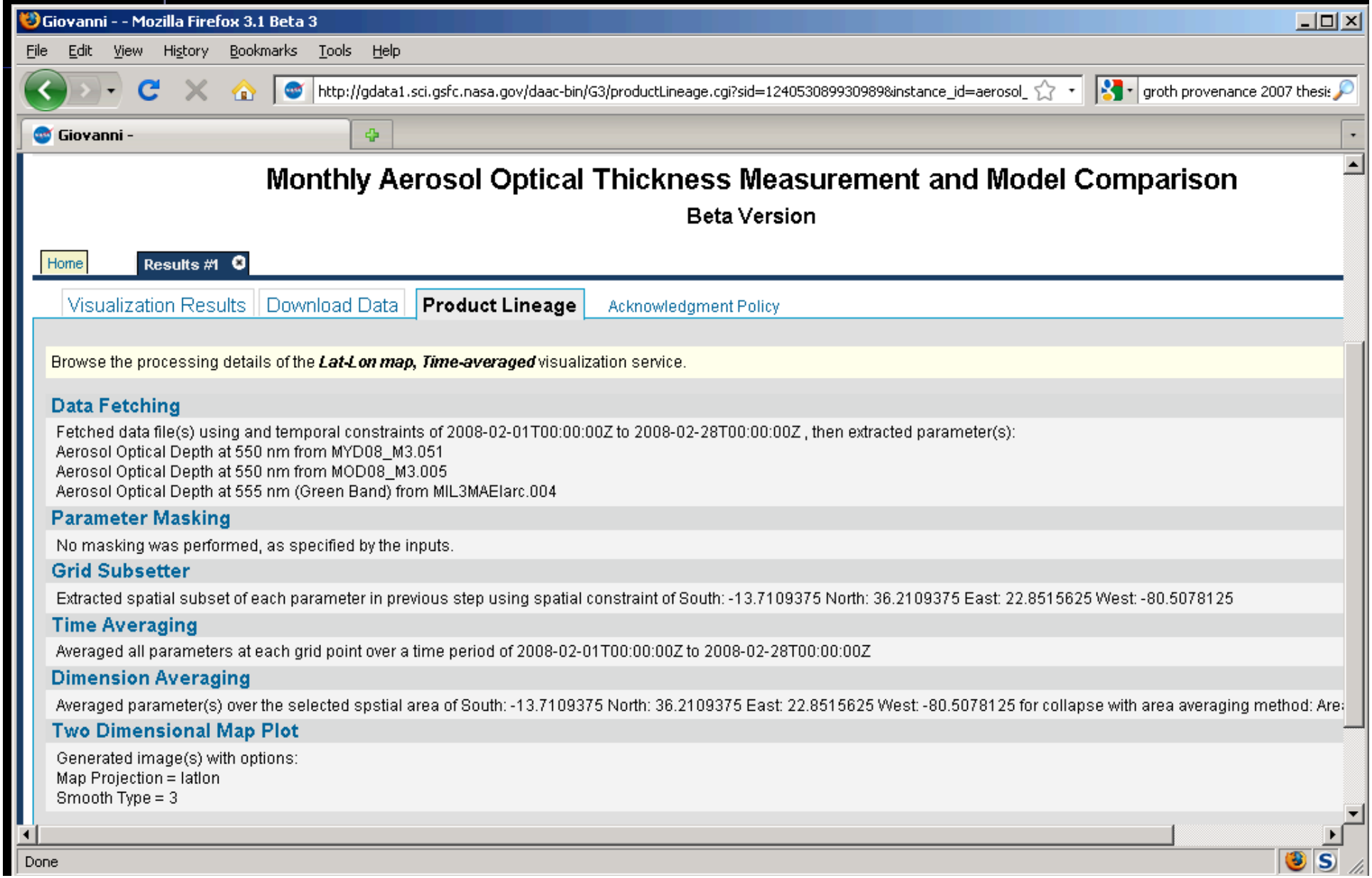
- Data is coming in faster, in greater volumes and outstripping our ability to perform adequate quality control
- Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision
- We often fail to capture, represent and propagate manually generated information that need to go with the data flows
- Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects
- The task of event determination and feature classification is onerous and we don't do it until after we get the data

A Problem of Confidence

- The ability to reproduce is the main factor in the confidence scientists have in a result.
- The ability to interpret and understand a result.
- The ability to understand the experiment and chain of reasoning that was used in the production of a result.
- The ability to verify that the experiment responsible for a result was performed according to acceptable procedures.
- The ability to identify what the inputs to an experiment were and where they came from.
- The ability to know who performed an experiment and who is responsible for its results.

From P. Groth, 2007

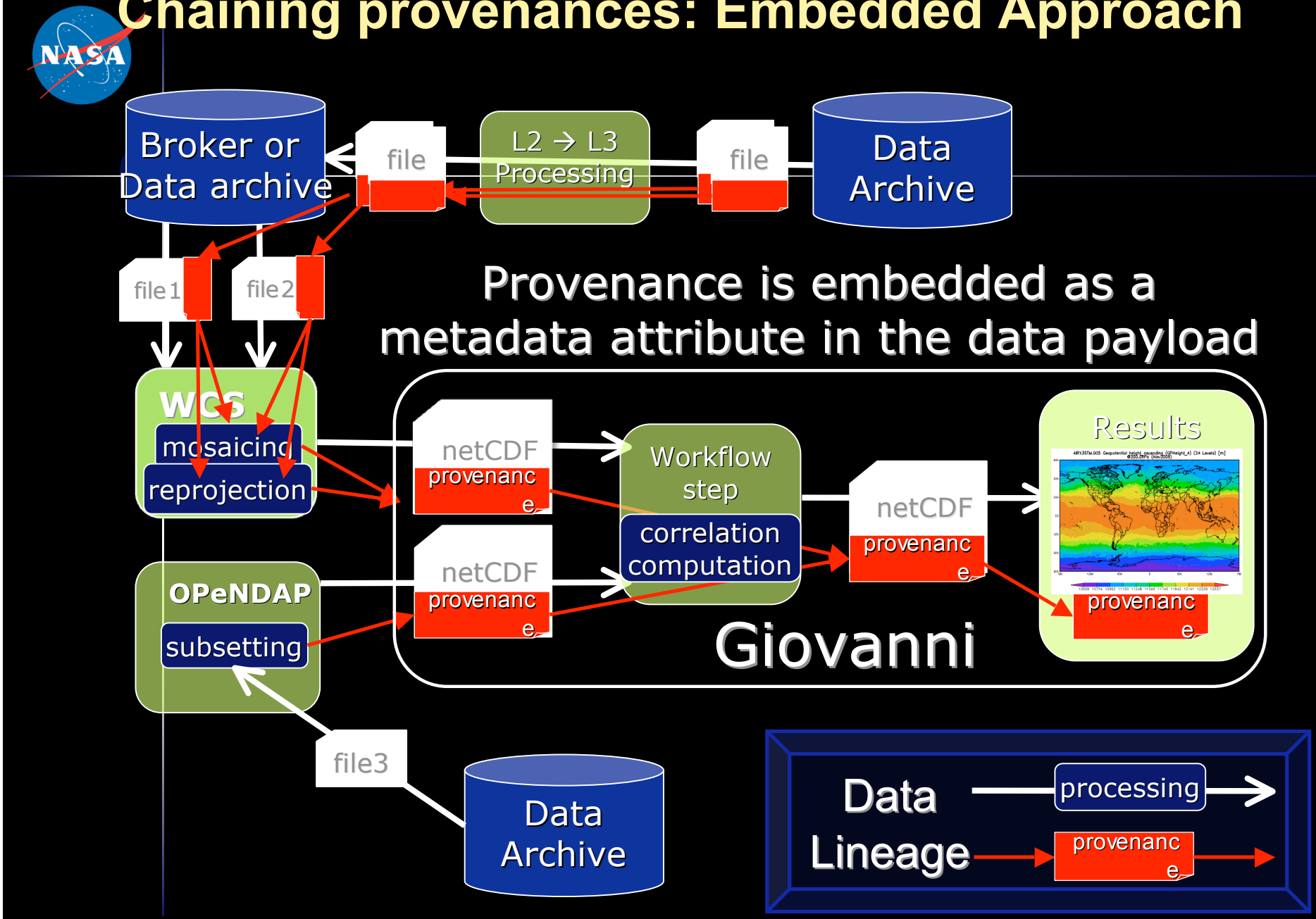
Product lineage in Giovanni



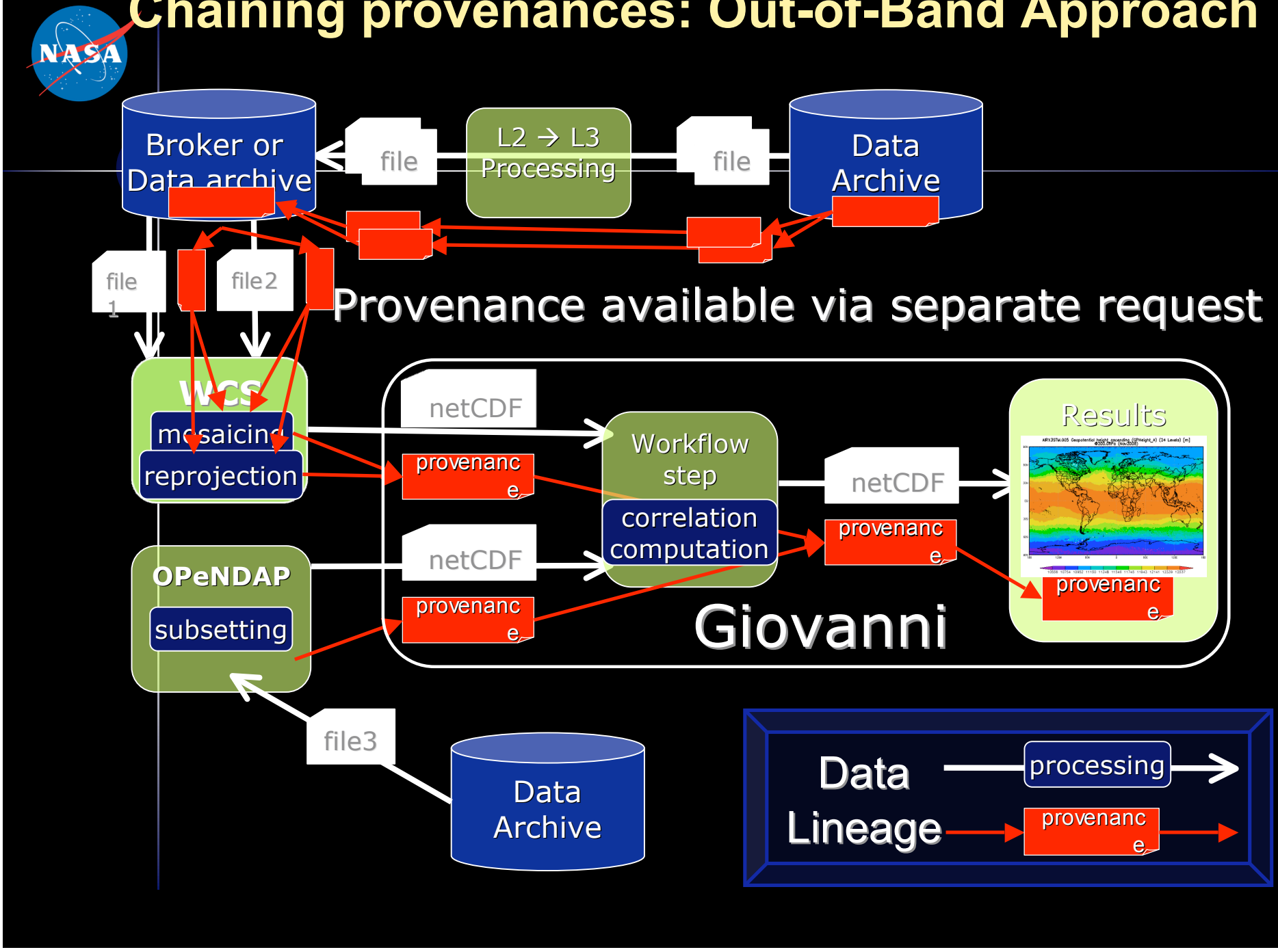
Provenance for Intercomparison

- *Automated or semi-automated intercomparison* of two apparently comparable parameters expose a challenge of the proper consideration of the data provenance.
- Dealing with two or more provenance chains is much more difficult.
- Provenance should be described with enough *semantic richness* to assess and eventually *assure the scientific validity* of an intercomparison operation.
- Complicating this task is the dispersion of data and services to multiple sources, to be accessed via *heterogeneous workflows*.
- Persisting and transmitting the rich provenance requires *provenance interoperability* in addition to data interoperability.

Chaining provenances: Embedded Approach



Chaining provenances: Out-of-Band Approach



How does the community want to use the data provenance?

- When (*i.e.*, under what conditions) do you pay attention to data provenance?
- What is the most useful data provenance information needed?
- How do you use the data provenance information?
- What other data provenance information would you like to see captured?
- How far should we go in automating data provenance capture, display, and utilization?